CTLS

OBJECTIVE TESTS

Janet G. Donald

Centre for Teaching and Learning Services
McGill University

OBJECTIVE TESTS

A Guide to Assist Faculty

prepared by
Janet G. Donald
Centre for Teaching and Learning Services

with the assistance of
Mary Bates and Joetta Browns

Published by University Teaching and Learning

McGill University 1978, revised 1983

Dépôt Légal
Bibliothèque Nationale
Premier Trimestre 1979

ISBN 0-7717-0042-3

TABLE OF CONTENTS

		Page
DEVELOPING GOOD OBJECTIVE TESTS		1
STEP 1	DEFINING WHAT YOU WANT TO TEST	3
STEP 2	CONSTRUCTING THE TEST	5
	Multiple Choice Test	5
	Matching Test	13
	True - False Test	18
	Short Answer or Completion Test	22
STEP 3	CHOOSING, ADMINISTERING, SCORING AND ANALYZING OBJECTIVE TESTS	27
REFERENCES		38

Preface

This booklet is a compact guide to the preparation of objective tests; it deals with the issues and problems encountered in these widely used methods of student evaluation.

The reader is cautioned that while objective tests are effective under appropriate circumstances, other kinds of tests such as essay, oral, and performance tests can also be effective methods of evaluating student learning. The references at the end of this booklet and the consultants at the Centre for Teaching and Learning Services are available to provide additional information and assistance in the preparation of objective tests and other evaluation methods.

The Editorial Group which supervised the preparation of this booklet consisted of Professors R.D. Japp, S.J. Shapiro, and P.C. Landry. Useful suggestions made by McGill faculty are gratefully acknowledged.

DEVELOPING GOOD OBJECTIVE TESTS

This booklet has been written to assist professors to construct and score objective tests. Objective tests consist of specific questions which require brief answers. They are called "objective" tests because equally competent scorers can score them independently and obtain the same results. The different types of objective tests are:

Multiple choice,
matching,
true-false, and
short answer or completion.

Objective tests have different advantages and disadvantages. The most important advantage of any objective test is that it permits reliable measurement of an extensive sample of course material. It can be adapted to a variety of teaching goals and subject areas; it is suitable for testing large groups and can be graded quickly and objectively. There are, however, several disadvantages of objective tests: preparation of test items can be time-consuming and costly; tests may over-emphasize rote learning and encourage guessing. Test item writers must guard against ambiguous wording, statements taken out of context, and a tendency to measure simple rather than complex kinds of learning.

Developing good objective tests requires three steps:

STEP 1 Defining what you want to test

- Define specifically what you want students to achieve in your course.

STEP 2 Constructing the test

- Write test items using the rules for constructing objective tests.

STEP 3 Choosing, administering, scoring and analyzing objective tests

- Choose among test formats.
- Administer and score test items.
- Analyze objective test items.

This booklet has been designed so that you can produce and evaluate your own test items. If you would like to test your knowledge of test construction, exercises and self-tests are provided throughout the booklet.

DEFINING WHAT YOU WANT TO TEST

STEP 1 Define specifically what you want students to achieve in your course.

This means,

1 You need to know not only your own subject matter but how students learn the material in your course. An analysis of your course content will enable you to produce a detailed course content outline. This will be an important aid in producing good test items.

2 You need a set of learning aims or goals. These will include learning facts, methods, and principles, making generalizations and interpretations, applying knowledge, and analyzing material. The principle is to be able to state what you want your students to know or to be able to do at the end of the course. Objective tests can be used to test the kinds of learning and learning outcomes listed in the following table.

Criteria and methods for producing a course content outline and a set of learning goals can be found in the references marked with an asterisk on page 38. Attention should be paid to the weight or importance assigned to different parts of the course content and to different learning goals.

KINDS OF LEARNING	EXAMPLES OF LEARNING OUTCOMES
KNOWLEDGE	
Knowledge of specific facts; terminology	Historical dates; math symbols; word meanings
Ways and means of dealing with these specifics	Chronological sequences; patterns of organization such as an hierarchical outline
Knowledge of conventions; classifications, criteria, methodology	Grammatical rules; classification system of plants; criteria for judging a musical performance; the steps in the scientific method of inquiry
Knowledge of principles, theories and structures	The laws of physics; the structure of the branches of government
COMPREHENSION	
Ability to "translate" knowledge or ideas	Changing a long poem into simple prose
Ability to identify and interpret the major idea in a communication or document	Explaining a section of the BNA act
Ability to extrapolate, i.e., to extend trends or project tendencies	Projecting the cost of living increase for the year
APPLICATION	
Ability to apply course knowledge to the solution of problems in an unfamiliar setting	Applying the various concepts of evolution to support or reject certain hereditary phenomena
ANALYSIS	
Ability to analyze a communication by distinguishing between fact and value statements	Distinguish between fact and unfounded assertion in an automobile advertisement
Ability to relate elements to some pattern of relationship or hypothesis	Distinguish between relevant and irrelevant facts in a debate and relate them to the total argument
Ability to infer the underlying principles of a communication	Determine the philosophy of an author from his writings

Adapted From Bloom (1956)

CONSTRUCTING THE TEST

STEP 2 Write test items using the rules for constructing objective tests.

MULTIPLE CHOICE TEST

The multiple choice test item consists of a stem, which presents a problem situation and several alternatives which provide possible solutions to the problem. The stem may consist of a question or an incomplete statement. The alternatives include either the correct or the best answer and several plausible but wrong answers, called distracters. The following two examples illustrate the question form and the incomplete statement form of multiple choice items.

Question form

Why do living organisms need oxygen?

a to purify the blood.
b to oxidize waste.
c to release energy.
d to assimilate food.
e to fight infection.

Incomplete statement

Living organisms need oxygen in order

a to purify the blood.
b to oxidize waste.
c to release energy.
d to assimilate food.
e to fight infection.

Particular advantages and disadvantages:

Multiple choice items can be used to measure most kinds of learning, but care and consideration is required to write good items. The items can and should be refined through item analysis and can be banked for future use (see STEP 3). Multiple choice tests require greater preparation time, but minimal scoring time compared with other test forms.

Exercise on defining what you want to test:

Write three statements of what you want your students to know. These statements should be derived from the learning goals in your course as discussed in STEP 1. For example,

> In dry air the percent by volume of oxygen is 21%.
>
> Determine the most cost effective way of heating and ventilating a house.
>
> Identify the principles critical to the solution of labor disputes.

You will use your statements as a basis for developing multiple choice test items.

RULES FOR CONSTRUCTION OF MULTIPLE CHOICE TEST ITEMS

These rules should be considered as guidelines or preferred practices.

RULE 1 Each item must be designed to measure achievement of the specific learning goal for which the student is being tested. Use the chart of Kinds of Learning and Learning Outcomes (p. 4) to guide your decisions. Caution: do not attempt to increase the difficulty of items by using more obscure and less significant learning goals or content areas.

RULE 2 A multiple choice test item presents a single definite statement or stem to be completed or answered by one of several given choices. The important concern is to ensure that the statement gives the student a clear understanding of what criterion is to be used in selecting the correct choice. An item such as the following encourages guessing and illustrates an "undefined" stem:

POOR ITEM:

To engage in international trade, a country must

a have a strong ruling class.
b have a large population.
c have a large labor supply.
d rise above subsistence living and develop surpluses.
e conquer neighboring countries.

By changing the stem to read "According to economic history, before a country becomes interested in trade it must first", attention is focused on the first prerequisite for trading. The task is now more clearly specified.

Exercise on constructing multiple choice test items:

For each of the three statements you wrote, construct a single multiple choice test item. Read through the remaining rules, returning to your three items as necessary, and if a rule has been violated, revise the item accordingly.

RULE 3 Keep student reading effort to a minimum by

- stating the item in simple, clear language
- putting as much of the wording as possible in the stem
- avoiding irrelevant material, and
- arranging response choices in logical order, e.g., alphabetically, numerically, or chronologically.

POOR ITEM:

Objective test items can be used to measure a variety of knowledge outcomes. The term "objective"

a refers to the method of identifying learning outcomes.
b refers to the method of presenting the problem.
c refers to the method of scoring the answers.
d refers to the method of selecting the test content.

In this example the same material is repeated unnecessarily in each alternative and the stem contains irrelevant material. The item is made much clearer and the student's reading effort is kept to a minimum if the following phrasing is used.

BETTER ITEM:

In objective testing, the term "objective" refers to the method of

a identifying learning outcomes.
b presenting the problem.
c selecting the test content.
d scoring the answers.

RULE 4 Ensure that the intended answer is the best choice. Avoid ambiguity. The authority should be quoted if the item contains controversial opinion, for the student may either disagree with or be unfamiliar with the instructor's opinion.

RULE 5 Avoid negative statements whenever possible; the negative multiple choice item confuses some students unnecessarily. Students often answer such items incorrectly even when they know the correct answers. If negative words must be used, capitalize, underscore, or otherwise highlight the negative word.

EXAMPLE

Which of the following is NOT a basic marketing management function?

a Advertising.
b Buying.
c Selling.
d Storing.
e Transporting.

This item assures that the negative wording will be noticed.

RULE 6 Use only plausible and attractive distracters as incorrect response choices. With some material, plausible incorrect alternatives are difficult to create. When it becomes impossible to construct a multiple choice item without including obviously incorrect alternatives, switch to the true-false or short answer test form. This eliminates the need for alternative responses.

RULE 7 Avoid giving clues to the correct choice.

·If possible, avoid using "a" or "an" at the end of the stem.

POOR ITEM:

The characteristic feature of a herpes virus is that it has an

a enveloped virion with icosahedral capsid.
b non-enveloped virion with icosahedral capsid.
c enveloped virion possessing no common antigens.
d enveloped brick-shaped virion.

BETTER ITEM:

Herpes viruses are characterized by having

a enveloped virions with icosahedral capsids.
b non-enveloped virions with icosahedral capsids.
c enveloped virions possessing no common antigens.
d enveloped, brick-shaped virions.

In cases where it is impossible to rephrase in this way, include both indefinite articles in the stem, as "a/an".

·The relative length of the correct answer may be a clue; ensure that all alternatives are of similar length.

·Avoid verbal clues which might enable students to select the correct answer or to eliminate an incorrect alternative. A key word in the stem may be unintentionally repeated in the correct answer in the form of a synonym, for example:

POOR ITEM:

Which of the following would you consult first to locate research articles on hereditary disorders?

a Journal of Abnormal Psychology.
b Journal of Genetic Psychology.
c Physiology and Behavior.
d Psychonomic Science.

The word "hereditary" in the stem points to the correct alternative which contains the word "genetic".

- Ensure that each complete multiple choice item is independent of other items on the test and gives no clue to the answers of the other items.

- The position of the correct responses should be varied randomly. Placement of the correct answer should not follow a pattern that may be apparent to a student taking the test.

RULE 8 Try to avoid using "all of the above" or "none of the above". Because these phrases are different in form from other distracters, they may make the choice of answer easier or confusing for students. In any case, they change the student's task.

RULE 9 Arranging the test items:

- items should be arranged so that all items of the same question type are grouped together. In a test that contains several different types of objective test items, a separate section should be allotted to each item type.

- items should be arranged in logical order, generally of increasing difficulty.

- for some purposes, it may be useful to group together items which measure the same learning outcomes or which cover the same content area.

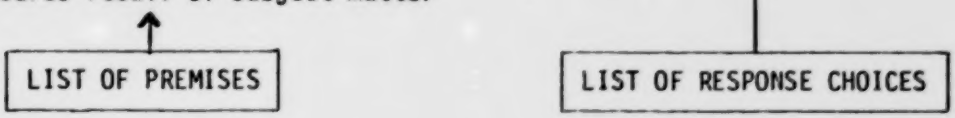
MATCHING TEST

The matching item is very similar in form to the multiple choice item. Two lists are presented side by side: on one side are listed a series of stems called premises, and on the other side are listed the alternatives or response choices.

EXAMPLE 1

Directions: on the line preceding each item in column I, put the letter of your choice from column II that best identifies the item. Each of the alternatives in column II may be used more than once.

Column I	Column II
_____ Best for measuring computational skills	a matching
_____ Least useful for education diagnosis	b short answer
_____ Measures greatest variety of learning outcomes	c multiple choice
_____ Most difficult to score objectively	d true-false
_____ Measures recall of subject matter	



Particular advantages and disadvantages:

Matching items are most useful for measuring recognition and recall. They can be used in most areas where knowledge of facts is important, and are useful for items which have several descriptive qualities. Matching items have limited use because all responses should be plausible alternatives for each premise in order to prevent guessing correct answers by the process of elimination. This condition is difficult to meet in many subject areas.

Matching test items measure the student's ability to identify the relationship between a premise and a response. You can construct items to test either the student's recall of knowledge or his or her ability to apply that knowledge. In testing for recall, as in example 1, the student is required to match terms (the list of response choices) with their characteristics (the list of premises). In example 2 the ability to apply knowledge is tested: a list of physical laws constitutes the response choices and the student is directed to apply the laws to novel situations as described in the list of premises.

EXAMPLE 2

Directions: for each description in column I select the letter of your choice of physical law from column II that best applies to it. Each physical law may be used more than once. Mark your answer on the accompanying answer sheet.

Column I	Column II
1 Equal force was applied to objects A and B. A had twice the mass of B. A traveled 10 km, B, 20 km.	a The law of action and reaction.
2 Two astronauts in deep space push against each other and move in opposite directions.	b The law of inertia
3 Once in motion, a car is easy to push.	c The law of motion.
4 If a force of 2 G's is applied to a cart, its speed will be 3 kmh, if 4 G's, 6 kmh.	

Exercise on constructing matching test items:

Using material from your course, write a set of premises and a corresponding list of response choices such that the relationship between the two requires recall of knowledge. Then write an introductory statement giving directions for answering the item. In the same way construct another matching test item, this time testing application of knowledge.

While reading through the rules for constructing satisfactory items, examine the items you have designed to ensure that they agree with each rule. Revise them where necessary.

RULES FOR CONSTRUCTION OF MATCHING TEST ITEMS

RULE 1 In the directions:

- specify the basis for matching premise and response.
- indicate whether response choices may be used more than once; permitting students to use the same response more than once usually reduces guessing.
- indicate where to write answers.

RULE 2 Select homogeneous premises and response choices. For example, if the item is intended to test knowledge of the location of European countries (names of countries listed as response choices), only the names of countries which are reasonable candidates should be included. Students will immediately eliminate implausible choices.

RULE 3 Keep the lists of entries short and place the brief responses on the right. Brief, succinct premises and responses save reading time, allowing the student to devote effort to the matching task itself.

RULE 4 Provide unequal numbers of response choices in relation to premises. This difference becomes critical in the case where choices can be used only once; here the correct answer may be guessed by simple elimination. To reduce this possibility, either (a) have more responses than premises, or (b) make some responses fit more than one premise.

RULE 5 Responses should be presented in logical order, that is, alphabetically, chronologically, etc. This makes the task clearer and more direct.

Self Test on Matching Test Items

To review your understanding of matching test items, read each of the following examples of matching items. If the example violates one or more of the stated guidelines, write down the number of the rule(s) (as on p. 16) violated. If the example does not violate any of the guidelines, put a check in the "OK" column. Answers are on the following page.

VIOLATED GUIDELINES | OK

1. Directions: Which item in Column I goes with each one in Column II?

	Column I		Column II
1	The carrier of yellow fever	a	Heart disease
2	The black plague	b	14th Century
3	Carriers of disease germs	c	Mosquito
4	The leading "killer" disease	d	Vectors

___ ___

2. Directions: Listed in Column I are mathematical formulae. Match the formula with the name of the phenomenon described under II. Remember the names may be used more than once. Be sure to mark your answers to the left of Column I.

	Column I		Column II
1	_____ $y = mx + c$	a	Circle
2	_____ $y^2 = 4fx$	b	Ellipse
3	_____ $\frac{x}{a}2 + \frac{y}{b}2 = 1$	c	Hyperbola
4	_____ $\frac{x^2}{a}2 - \frac{y^2}{b}2 = 1$	d	Parabola
		e	Straight line

___ ___

3. Directions: Select one best answer for each item in Column I from the statements in Column II.

	Column I		Column II
1	_____ Szasz	a	Emphasis on socio-cultural causes of mental disorder
2	_____ Horney	b	Wild outbursts of homicidal aggression
3	_____ Epidemiology	c	Incidence and distribution of disease or disorder
4	_____ Amok	d	"Illness" as an inappropriate term for abnormal behavior

___ ___

TRUE - FALSE TEST

The true-false test item generally consists of a declarative statement which is to be judged true or false as in examples 1 and 2.

EXAMPLE 1

T F The cost of goods sold is found by adding the net cost of delivered purchases to the beginning inventory at cost.

EXAMPLE 2

The volume of a mass of gas:

T F a tends to increase as the temperature increases.
T F b tends to increase as the pressure increases.
T F c may be held constant by increasing pressure and decreasing temperature.
T F d may be reduced by increasing pressure and decreasing temperature.

True-false items can be further developed by adding fill-in responses as in example 3. To answer these items, the student must analyze false statements. This increases the possibility of measuring students' comprehension of the statement and eliminates guessing to a great extent.

EXAMPLE 3

Directions: make every false statement true by suggesting or substituting for the underlined words(s).

The use of steam revolutionized transportation in the seventeenth century. TRUE FALSE ( )

Operating ratio may be defined as the funds available to cover the expenses of operating a business after the cost of goods sold is subtracted from net sales. TRUE FALSE (Gross Margin)

Answers to Self-Test on Matching Test Items: Rules broken were:

1. 1, 2, 4, 5 2. OK 3. 1, 2, 3, 4, 5

Particular advantages and disadvantages:

Because of the ease of recording a response, a broad range of material can be evaluated quickly. Since there is a 50-50 chance that a student can choose the correct answer, the true-false item encourages guessing. By asking students to correct false statements as in example 3, guessing can be eliminated to a certain extent and more diagnostic information is revealed. Test scores can be adjusted to compensate for guessing as described in the section on Scoring the True-False Test (p. 29). The author must be on guard against statements taken out of context or with too narrow a focus. At the same time, one must also guard against statements which are overgeneralizations and cannot therefore be absolutely confirmed or denied.

Exercise on constructing True-False test items:

Using material from your course, write three factual statements that could form the basis of true-false test items. For each statement construct a test item that consists of only two choices as in example 1 and 2 and an item that includes fill-in responses as in example 3. Read through the rules for constructing good true-false tests and check your items to ensure that they agree with each rule. Revise them if necessary.

RULES FOR CONSTRUCTION OF TRUE-FALSE TEST ITEMS

RULE 1 Each item must be designed to measure achievement of a single learning outcome. Furthermore, trivial details should not make a statement true or false; a statement should, instead, be a fundamentally true or false idea.

POOR ITEM:

T F The XYZ Intelligence Test contains 88 items divided equally among 8 subtests.

BETTER ITEM:

T F The XYZ Intelligence Test has been validated as being "culturally unbiased".

Here, the better item has more worthwhile content; knowledge regarding the validity of an intelligence test is important, whereas remembering the exact number of its items and subtests is of minor importance.

RULE 2 The statement should be phrased in such a way that it can be judged categorically true or false. To ensure this:

- avoid broad, general statements and indefinite or ambiguous terms such as rarely, several, may, possible.
- include only one central significant idea in each statement.
- keep the statements short and use simple language structure.
- use negative statements sparingly; if they must be used, highlight the negative term. Avoid two negatives in the same sentence.
- attribute statements of opinion to a source or authority.

RULE 3 Avoid unintentional clues to the answer.

- avoid length clues by making all statements of similar length.

- avoid using terms which indicate the answer. For example, terms such as always, never, every, none, all are likely to indicate the answer false or no, while terms such as often, generally, usually, maybe, point toward true or yes responses.

- avoid patterns of responses by randomly sequencing responses and ensure that there is an approximately even proportion of true and false responses.

SHORT ANSWER TEST

Test items that require the student to give specific information in his or her own words are classed as short answer or completion items. They may be presented as questions, incomplete statements, definitions, identifications, or analogies.

Question form

What is the term which Freud uses to describe personal energy?

_____

Incomplete statement

The kinds of learning that objective tests can be used for include:

_____

Definition

Define "test" _____

_____

Identification

Identify or list the forms of tests recommended for measuring comprehension.

_____

_____

Analogy

Mendel is to biology as Einstein is to _____

Particular advantages and disadvantages

Short answer items are relatively fast and easy to construct. The short answer test also has the advantage over other kinds of objective tests in that the student must construct an answer rather than just recognize a response. Analogies can be used to test students' ability to analyze. It is, however, very difficult to phrase items so that only one answer is correct; usually more than one word can be used as a response and this affects the ease and potential objectivity of scoring.

Exercise on constructing short answer test items:

Using material from your course, write a short answer test item for each form. Read through the rules for constructing good short answer tests and check your items to ensure that they agree with each rule. Revise them where necessary.

RULES FOR CONSTRUCTION OF SHORT ANSWER ITEMS

RULE 1 As far as possible, state the item so that only a single, brief answer can be given. The stem of the item should convey to the student the limits of responses. For example, the following item does not clearly specify how much information is required for a correct answer.

POOR ITEM:

What are the functions of the mitochondrion?

BETTER ITEM:

What are the three main functions of the mitochondrion?

In the better item the task is well defined and scoring is rendered more objective since each correct function listed can be counted as one point.

RULE 2 A direct question, rather than an incomplete statement, is more likely to present the problem clearly and allow for only one answer. If you wish to employ an incomplete statement, first construct it as a question, and then change it to an incomplete statement.

POOR ITEM:

Evaluation can be used for _____

_____

BETTER ITEM:

(What are five uses of evaluation?) Five uses of evaluation are _____

_____

RULE 3 The words to be supplied for the response should relate to the main point of the statement; the omitted part must be a key word or phrase.

EXAMPLE:

Emmerts' Law states that the apparent (size) of an after-image is directly related to its apparent (distance).

Omitting the words "size" and "distance" tests the students' knowledge of what factors are related in Emmerts' Law, whereas omitting the words "law" or "after-image" would give negligible information as to the students' grasp of the law.

RULE 4 Wherever possible, the blank space(s) should be placed at or near the end of the statement. This permits the student to read the complete problem before coming to the blank and it minimizes reading effort.

RULE 5 Avoid extraneous clues to the answer.

·Grammatical clues must be avoided. For example,

POOR ITEM:

> An engine that uses explosive gas as its energy source is called an (internal combustion) engine.

The word "an" before the blank dictates that the answer must begin with a vowel. This can be rectified by substituting "a/an", or better still, the item could be rephrased as follows:

BETTER ITEM:

> What do you call an engine that uses explosive gas as its energy source?

·Different lengths of answer blanks are an obvious clue to likely responses. Ensure that all blanks are of the same length, and allow sufficient space for the longest response.

RULE 6 For numerical answers, indicate the degree of precision expected and the units in which the answer is to be expressed.

EXAMPLE:

What is the distance in kilometres from London to Rome?

_____

STEP 3 CHOOSING, ADMINISTERING, SCORING AND ANALYSING OBJECTIVE TESTS

How to choose among test formats

The first criterion to use is the type of learning outcome you are measuring. All objective test forms are adequate for testing knowledge. Comprehension, application and analysis skills may be better tested by multiple choice and short answer items than by true-false or matching items. Where sufficient alternatives exist, multiple choice and matching items are preferred. Least guessing could be expected with short answer items and most with true-false and matching items. Multiple choice items are the most difficult to produce but can be analyzed and used repeatedly. Multiple choice, true-false, and matching tests can be machine scored.

Gronlund (1977) suggests that the item author should start each item as a multiple-choice item, switching to another item type only when the learning outcome or subject matter makes it desirable to do so. Thus, when there are only two possible alternatives, a shift can be made to a true-false item. When there are a number of similar factors to be related, a matching item would be preferable, and when recall or the ability to supply an answer is important, a short-answer item should be used.

Considerations for administering objective tests

·The professor, because of familiarity with the language of his or her own discipline, might not recognize ambiguities and other flaws in test items. The following precautionary measure should be taken: newly-composed objective test questions should be submitted to experienced colleagues for thorough appraisal. Also, trying items out with a few students is a good way of pointing out ambiguities before the test is used. Former students make ideal appraisers.

·No matter how well prepared the test is, there is always a chance that a student will find a question that can be given a quite unintended interpretation. In some departments, students are allowed to explain their interpretations in a written challenge. The following kind of statement is included in the test instructions,

> "If you wish to challenge a question because you think it unclear or ambiguous, note the number of the question here (on a signed check sheet to be returned along with the answer sheet) and write a note to explain how you have interpreted it."

Students should know in advance that a challenge, depending on the degree of insight it demonstrates, may lead the examiner to accept a "wrong" answer. If the examiner acts on challenges before the answer sheets go for computation, this feature is likely to take only an hour or so, even for a class of several hundred.

·To avoid cheating, several versions of the test paper could be prepared with the questions shuffled into different sequences. In each version of the test paper, items should still be ordered such that the easier ones appear before the more difficult, and items dealing with the same subject matter are grouped together.

Scoring multiple choice and true-false tests

In multiple choice and true-false tests, students have a chance to guess the correct answer. Particularly in true-false or speed tests, a more conservative estimate of student achievement can be obtained by use of a correction formula. If the correction for guessing is to be used, students should be informed of this before they take the test.

$$\text{Formula: SCORE} = \text{RIGHTS} - \frac{\text{WRONGS}}{n-1} \quad \text{where}$$

n is the number of choices in the item. Thus for a 40 item true-false test on which a student answers 30 items correctly and 5 incorrectly, and omits 5 items,

$$\text{the score} = 30 - \frac{5}{(2-1)} = 25$$

Scoring short answer tests

Since short answer and completion items require the student to construct his or her own responses, these items cannot be scored with complete objectivity. It is therefore a good idea to compile a list of the possible correct answers and then to submit the test to a colleague for thorough appraisal. Other possible right answers may thus come to light and unforeseen flaws may be pointed out. Even with this precaution, still other acceptable answers may be offered by students writing the test; if these answers are considered right, they should be added to the list for future reference.

Analyzing Objective Test Items

Item analysis is a procedure for studying the students' responses to test items. It indicates to the professor those items which require revision or improvement and gives the student a clear indication of where his or her mistakes lie. An item analysis will provide information concerning: the difficulty of the item; the discriminating power of the item; the plausibility of the incorrect response choices.

The McGill Computing Centre offers an exam scoring service that provides, along with the students' scores, the statistics for analyzing your test items.

1 Item difficulty: How difficult is the item? How many students pass the item? Item difficulty is reported as a percentage of students getting the item correct. Where the instructor is devising items that will differentiate among students, an item difficulty of 50%, with an acceptable range of between 25% to 75%, is the standard for choosing an item. In a selection test, then, we would aim to have 50% of the students answering the item correctly, but accept items which are answered correctly by 25-75% of the students. On a mastery test, however, a more important criterion for choosing a test item is how well it matches the learning priorities in the course. The asterisked references on page 33 deal with this question in greater detail.

2. Item discrimination: Does the item accurately separate students into low and high achievers? Do the "good" students get the item right and the "poor" students get the item wrong? The McGill exam scoring service provides a Point-Biserial correlation in which individual student's scores on each item are compared with their overall test scores. As a result, each test item has a point-biserial correlation which is a measure of the strength of relationship between correct answers to the item and a high score on the overall test.

The correlation can range in value from -1.00 to +1.00. However, it is extremely rare for an item to have a point-biserial correlation greater than 0.80.

A correlation between 0.0 and 0.2 represents an item that is unable to accurately test students' ability. Values in the range 0.3 to 0.4 identify good items. Greater than 0.5, the item is very difficult and only the top students can get it right.

A negative correlation generally indicates an ambiguous item. In this case, a large portion of good students have selected a distracter rather than the correct answer. This situation may also occur when the wrong answer number has been provided as the correct answer during scoring.

The Point-Biserial correlation is only a statistic, and requires a reasonable sample size to make the above ranges dependable. Tests administered to classes of less than 50 students cause the correlation to less distinctly gauge the power of a question. In this case, more attention should be paid to the item difficulty and plausibility for accurate analysis.

Plausibility of incorrect response choices or distracters on multiple choice tests

Compare the number of students in the upper and lower 25% who selected each incorrect alternative. A good distracter will attract more students from the lower 25% than from the upper 25%.

EXAMPLE For a class of 52 students,

Response Choices	A	B*	C	D
Upper 25%	1	6	6	0
Lower 25%	5	2	6	0

*correct response

Here, response choice A is a good, plausible distracter, C is poor and D is ineffective. The McGill exam scoring service provides a listing of student responses from which distracter plausibility can be derived.

Items that discriminate adequately but have poor distracters can be re-written for future use. Item analyses should be stored with the test items for future reference. Green (1975) provides a more detailed method of item analysis.

Marking of Multiple Choice Examinations by Computer through the Computer Centre

The Computing Centre provides a service to score multiple choice examinations. Such examinations have the answers marked with a pencil on special cards, which are readable by computer, and reports and statistics are generated. There are several different computer programs available to produce these reports and the one used depends on how the exam is set up. In most, a maximum of 180 questions can be scored, but a new program that can score up to 540 questions is now available.

When the cards are received at the Computing Centre, they are first read through the card reader and placed on a disk data set. At the same time the cards are checked for blanks or double marks and the student number is checked to ensure that it falls within a valid range. Error messages are printed out when errors are detected and those cards are manually checked and the disk data set edited if necessary. The cards which have been changed in the edit step are turned over in the deck so that the results can be rechecked by the instructor. When all computer detected errors have been checked and corrected the appropriate control statements are added and the exam report is produced.

In this step the computer program matches the student number marked on the front of the card with a student number from a disk file which contains student numbers and names for all students registered at the university. If the numbers match the name is extracted and placed on

the report. If the numbers do not match the name will come out as blank. If the student has marked the wrong student number the results could be printed with another name (the one corresponding to the number).

Where to submit and collect processed results

Cards from examinations should be brought to the reception area in BH200 as soon as possible after the examination. Master card(s) with the correct answers should be brought at the same time as well as instructions on how the exam should be marked. A receipt is issued at that time. When the results are ready the instructor is called. The results are picked up from reception in BH200 upon presenting the receipt.

Turnaround time

During regular term time the turnaround times for class tests is a maximum of 2 working days.

During exam periods (midterms or finals) the exams are processed in the order they are received. Turnaround time is usually 1 to 3 working days but could be longer.

Instructor responsibility

1. Make sure that the right cards are available for student use (e.g. card two can not be used for questions 1-60).

2. Make sure that markings on cards are sufficiently dark and that erased marks are erased properly.
3. Provide properly marked master card(s).
4. Ensure that each card has the correct student number marked on the front of the card.
5. Remove all extraneous pencil marks. (Note that the backside of the card is not read by the computer and can be used for indication of Section, Row Number, etc.)

When the results are obtained:

6. Check the statistics to determine if a rerun is necessary with a changed master card or questions deleted.
7. Requests for reruns must be in writing and resubmitted to BH200
8. It is not the responsibility of the Computing Centre to determine if questions should be deleted.
9. Make sure that names appearing on the report appear on a classlist. If not determine where the error is and make sure that the mark is given to the right student.

Computing Centre responsibility

1. Process the student cards using the master card provided by the instructor to generate the standard computer output.
2. Ensure that blank or double responses to questions are correctly processed.

Charges

There is no charge for the standard service when it is used for McGill credit courses.

Charges will be made to cover the cost of extra personnel time required if sub-standard cards are submitted and/or additional runs are necessary. Such jobs will also result in additional service delays.

Card availability

The special cards used for examinations are available through the Computing Centre. The cost is $10.00/1000. For final examinations these cards are made available through the faculties but for class tests they are bought from BH200. There are three different cards available: card one has question number 1-60, card two question number 61-120 and card three question number 121-180.

Computer programs and results

All of the scoring programs provide: a list of student I.D. numbers and total scores suitable for posting; a separate list of student names, I.D. numbers, scores by item, and total scores; frequency distribution of total scores; class mean and standard deviation; answer distribution and Point-Biserial correlation for each item. The different scoring programs are:

1. The NORMAL program provides for separate integer weights for each item plus a penalty for incorrect or missing answers.

2. The WEIGHTED program allows for separate weights to be assigned to each answer alternative of every item. This program may be used to score examinations when an item is discovered to have more than one correct answer.

3. The MULTIPLE program scores up to 9 differently sequenced versions of the same examination. Answer weighting is as for the NORMAL program.

4. An exam randomizing program is available to scramble an examination into 2 to 9 differently sequenced versions. This limits the opportunity for student cheating.

5. The MATCHUP program has been specifically designed for multiple choice items which standardly have more than one correct answer. This program allows for separate weights for a student selecting a correct answer, selecting an incorrect answer, or missing (not selecting) a correct answer. All of the statistics provided by the other scoring programs are also available except for the Point-Biserial correlation, which is replaced by the Pearson correlation.

6. The above mentioned programs all mark a maximum of 180 questions. Another program is available which marks up to 540 questions. This program replaces the programs described under 1, 2, and 3. More statistics are calculated. The additional statistics are first quartile, median, and third qartile. In the item analysis Biserial correlations and a Reliability Index are also calculated.

   This program is used only when requested because of slower turnaround due to the program size.

   For more information please contact Mrs. Manley at 392-4845.

References

Bloom, B. S. (Ed.). Taxonomy of educational objectives, handbook I: Cognitive domain. New York: McKay, 1956.

*Bloom, B. S., Hastings, J. T., & Madaus, G. F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

*Brown, F. G. Measurement and evaluation. Itasca, Ill.: F. E. Peacock Publishers Inc., 1971.

Cranton, P. Constructing tests, module 6. Centre for Learning and Development, McGill University, Montreal, 1978.

Donald, J. Guide to methods of evaluation in higher education. Centre for Learning and Development, McGill University, Montreal, 1977.

Educational Testing Service. Making the classroom test . Princeton, 1969.

Gorow, F. F. Better classroom testing. San Francisco: Chandler Publishing Co., 1966.

*Green, J. A. Teacher-made tests (2nd ed.). New York: Harper and Row Publishers, 1975.

Gronlund, N. E. Constructing achievement tests. Englewood Cliffs, N.J.: Prentice-Hall Inc., 1977.

McBeath, R. J., et al. Faculty self-appraisal and development project. San Jose State University, 1976.

Milton, O., & Edgerley, J. The testing and grading of students. New York: Change Publication, 1976.